Research Article

A single-copy IS5-like transposon in the genome of a bdelloid rotifer

Eugene A. Gladyshev[1] and Irina R. Arkhipova[2*]

[1]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA; [2]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA.

*To whom correspondence should be addressed. E-mail: iarkhipova@mbl.edu.

Dr. Irina R. Arkhipova

Josephine Bay Paul Center for Comparative Molecular Biology and Evolution

Marine Biological Laboratory

7 MBL Street

Woods Hole, MA 02543

Tel. (508) 289-7120

Fax: (508) 457-4727

Key words: Lateral/horizontal gene transfer; DNA transposons; host factors; *Adineta vaga*.

Running head: Rotifer IS5-like transposon

Abbreviations: aa, amino acids, CD, conserved domain; HGT, horizontal gene transfer; IS, insertion sequence; ORF, open reading frame; PCR, polymerase chain reaction; TE, transposable element; TIR, terminal inverted repeat; TSD, target site duplication; UTR, untranslated region.

Abstract

In the course of sequencing telomeric chromosomal regions of the bdelloid rotifer *Adineta vaga*, we encountered an unusual DNA transposon. Unlike other bdelloid and, more generally, eukaryotic transposable elements (TEs), it exhibits similarity to prokaryotic insertion sequences (IS). Phylogenetic analysis indicates that this transposon, named IS5_Av, is related to the ISL2 group of the IS5 family of bacterial IS elements. Despite the apparent intactness of the single open reading frame coding for a DDE transposase and the perfect identity of its 213-bp terminal inverted repeats (TIRs), the element is present in only one copy per diploid genome. It does not exhibit any detectable levels of transcription, so that its transposase gene appears to be silent in the bdelloid host. While horizontal transfers of TEs between kingdoms are not known to happen in nature, it appears likely that IS5_Av underwent integration into the *A. vaga* genome relatively recently, but was not successful in adapting to the new host and failed to increase in copy number. Alternatively, it might be the only known member of a novel eukaryotic DNA TE superfamily which is so rare that its other members, if any, have not yet been identified in eukaryotic genomes sequenced to date.

Introduction

Transposable elements (TEs) are omnipresent in all three domains of life: Bacteria, Archaea, and Eukarya. Of the two major types of TEs, *i.e*. retrotransposons and DNA transposons, the latter are particularly prone to horizontal transmission (see Silva *et al*. 2004 for review). First inferred from patchy distribution of P-elements in Drosophila *spp*. (Kidwell 1983; Daniels *et al*. 1990) and mariner transposons in insects (Robertson 1993), horizontal gene transfer (HGT) was subsequently suggested to constitute the primary mode of survival for DNA transposons of this type (Robertson and Lampe 1995; Hartl *et al*. 1997). According to this scenario, a DNA TE enters a new host and proliferates within its genome, giving rise to multiple

copies, which then undergo silencing, mutational decay, and eventually get erased from the host genome, so that the long-term persistence of a TE depends on the ability of a functional copy to escape into a new host. During such horizontal escapes, however, TEs do not typically cross domain boundaries: no cases of recent HGT between Bacteria (or Archaea) and Eukarya have been documented to date, and naturally occurring cases incompatible with vertical inheritance typically involve movements between taxonomically close groups (Daniels *et al*. 1990; Diao *et al*. 2006; Pace *et al*. 2008; reviewed in Feschotte and Pritham 2007). Interkingdom transfers, *e.g*. from animals into protists and bacteria, have been accomplished in the laboratory (Gueiros-Filho and Beverley 1997; Rubin *et al*. 1999). In these cases, however, expression of the transposase gene involved its placement under a heterologous promoter known to be functional in the new host (*e.g*. Rubin *et al*. 1999; Zhang *et al*. 2000). Fundamental differences in prokaryotic and eukaryotic gene expression, such as incompatible basal promoter elements, translation start sites, or coupling between transcription and translation, may, at least in part, account for the barriers to successful interdomain transfers.

Here we describe an apparent case of horizontal transfer of an IS5-like DNA transposon, which was found during cloning and sequencing of telomeric regions in the bdelloid rotifer *Adineta vaga* (Gladyshev *et al*. 2008). As telomeric regions of bdelloid genomes contain numerous foreign genes originated in bacteria, fungi, or plants, it was therefore not as surprising as it could have been in the absence of massive HGT that a DNA transposon of apparently bacterial origin was also able to find its way into the bdelloid genome. Interestingly, however, this TE appears to have been much less successful in adapting to the new host environment than certain bacterial protein-coding genes, such as the D-Ala-D-Ala ligase gene which is transcribed, spliced, and encodes a functional protein (Gladyshev *et al*. 2008). As described below, IS5_Av integration apparently represented a one-time event and did not result in proliferation in the new host, possibly due to the lack of compatibility with the transcriptional/ translational apparatus, or damage to the element. Our study indicates that in nature, given the

opportunity, TEs highly similar to prokaryotic could incorporate into metazoan chromosomes. Alternatively, IS5_Av could represent the sole member of a hitherto unknown eukaryotic TE superfamily.

Materials and Methods

RT-PCR and Southern blot hybridization experiments were done as described in Gladyshev and Arkhipova (2007). The following primers were used for PCR amplification: TIR, GAGTAAAGTTTTGTGTTCACTG; RT-F1, GTCAGAATGGAGTCGCAACA; RT-R1, TCGGTGGTTACAACAATCACA. AvDdl primers were as in Galdyshev *et al*. (2008). Southern blots and genomic library were probed with the ^{32}P-labeled 1637-bp fragment amplified by the TIR primer, and with the 534-bp fragment amplified with the RT-F1/RT-R1 primer pair. The IS5 dataset was assembled from (i) IS Finder database entries (www-is.biotoul.fr), including at least 3 representatives from each subgroup of the IS5 group, and all representatives from the ISL2 group; (ii) Repbase entries (www.girinst.org/repbase/index.html), including representatives of the Harbinger/PIF and IS4EU/ISL2EU superfamilies; (iii) top GenBank hits from diverse archaea and bacteria identified by BLAST search; and (iv) IS5-like families in protist genomes identified by BLAST search. Amino acid sequences were aligned with T-Coffee (Notredame *et al*. 2000) and ClustalW implemented in MEGA4 (Tamura *et al*. 2007). Neighbor-joining and minimum evolution analyses were done with MEGA4 (p-distance or Poisson correction; pairwise deletion; 1000 bootstrap replications). The value of the parameter gamma was determined by ProtTest (Abascal *et al*. 2005). Alternative start codons were analyzed using the EasyGene server (www.cbs.dtu.dk/services/EasyGene/). Search for the TSD-TIR…TIR-TSD combination was performed on a 1419-bp region upstream and a 784-bp region downstream from the *hobo* ORF (between the stop codon of the upstream NHL gene and the 5' LTR of a downstream TE), using a custom Perl script allowing one or two mismatches. Only cases with 8-bp target site duplication (TSD) were considered, as *hobo*_Av is phylogenetically close to insect *hobo*

elements, and therefore should not deviate from the general rule specifying 8-bp TSDs for *hobo*-like TEs (Rubin *et al*. 2001). Unusual *hAT* elements yielding atypical 5-bp and 6-bp TSD (Putnam *et al*. 2007) belong to very distant and well-separated clades (data not shown).

Results

Structure and copy number

Structural organization of the *A. vaga* IS5-like transposon is shown in Fig. 1A. The transposon harbors a single ORF exhibiting similarity to transposases of the IS4/IS5 families (Table 1; pfam01609:Transposase_11, which includes prokaryotic transposases for IS4, IS421, IS5377, IS427, IS402, IS1355, and IS5, and is a member of the DDE megafamily of transposases/integrases), and which contains all of the highly conserved residues required for catalytic activity of these enzymes. A multiple sequence alignment of IS5-like transposases is provided as Supplementary data. The IS5_Av transposon is flanked by 213-bp perfect terminal inverted repeats (TIRs), including a 2-6 bp ambiguity in the outermost nucleotides CATATG… CATATG, all or part of which could also be regarded as a 2-, 4-, or 6-bp TSD (TA, ATAT, or CATATG). Such ambiguity can usually be resolved by comparing different flanking sequences from additional genomic TE copies. To our surprise, we were unable to identify any such copies, neither by exhaustive genomic library screens (data not shown), nor in Southern analyses using two different restriction enzyme combinations (Fig. 2a). The single band on a Southern blot of *A. vaga* genomic DNA digested with *Pvu*II, which has no recognition sites within IS5_Av, demonstrates that this is the only copy present in the genome. The size of the band (7.0 kb) coincides with the expected size of the *Pvu*II fragment harboring IS5_Av on the sequenced telomeric fosmid, ruling out possible contamination or misassembly in the course of random shotgun sequencing of fosmid subclones. Similarly, *Xho*I/*Hpa*I digest of genomic DNA yields an expected 4-kb band. In addition, sequencing of cloned and total PCR products obtained with the TIR primer, as expected, revealed no nucleotide sequence polymorphisms, which might have

been observed if different full-length copies existed in the genome but were not present in the genomic library (data not shown).

Genomic environment

IS5_Av is located in a subterminal region of the *A. vaga* telomere K_A (Gladyshev *et al*. 2008), in an environment rich in eukaryotic TEs that are typical of bdelloid rotifers (Arkhipova and Meselson 2005; Gladyshev *et al*. 2007). This particular region has no homologous partner among *A. vaga* chromosomes, since colinearity between homologous telomeres K_A and K_B begins ~50 kb proximally to IS5_Av (see Fig. S1 in Gladyshev *et al*. 2008), and therefore IS5_Av exists as the only copy per diploid genome. In Fig. 1B, it may be seen that this element is located in an exceptionally TE-rich environment: it is apparently inserted into the 3' UTR of a *hobo*-like transposon and is surrounded by three *mariner*-like, one retrovirus-like, and two *piggyBac*-like transposons. The distance from IS5_Av to the chromosome end is 19.6 kb. The *hobo* element has several defects in its open reading frame, and therefore likely represents an ancient insertion; other TEs also carry in-frame stop codons, frameshifts, or indels (Table 2). Interestingly, each *mariner* or *piggyBac* copy belongs to a different subfamily within the corresponding superfamily, underscoring the considerable diversity of DNA TEs in bdelloids.

In the absence of other copies similar in sequence to the *hobo* element in Fig. 1, we sought to identify its exact boundaries by scanning the immediate flanks (the regions between the *hobo* ORF and the next adjacent element) for the presence of an 8-bp TSD combined with a short inverted repeat sequence. The scanned region contained only one potential TSD-TIR-TIR-TSD combination (8-bp TSD+7-bp TIRs with 1 mismatch, aattgattAAATAAT...ATcATTTaattgatt) (Fig. 1b). This boundary yields a *hobo* ORF framed by *ca*. 600-bp UTRs, which is very similar to the UTR length in other known rotifer *hobo* families (Arkhipova and Meselson 2005 and unpublished), thereby placing IS5_Av into the *hobo* 3' UTR within 108 bp from its ORF end. Alternatively, one of the TIRs could have undergone deletion, in which case the boundaries would be difficult to define.

One could also consider a scenario under which the 213-bp TIRs belong to a different non-autonomous foldback-like element, with IS5_Av subsequently inserted between these TIRs. We searched for an additional, shorter internal TIR-TSD combination and could not find any putative TIRs longer than 10 bp (GCGACTGAAT…ATTCAGTCGC; Fig. 1A); if these internal TIRs are the true TIRs of IS5_Av, they are unusually short and are not surrounded by TSD.

Phylogenetic placement

In contrast to the neighboring elements (and all other TEs previously identified in bdelloid rotifers), which are typical of eukaryotes and yield significant BLAST hits to other eukaryotic TEs from the corresponding superfamilies (hAT, piggyBac, or mariner/Tc; Table 2), IS5_Av appears quite different: its top database hits come from bacteria, and not from eukaryotes (Tables 1,2). The putatively bacterial origin indicated by similarity scores is also suggested by phylogenetic analysis (Fig. 3). It may be seen that IS5_Av from *A. vaga* clusters with a specific subgroup of IS5-like TEs, namely the ISL2 group, which is characterized by 15-40 bp TIRs, and TSD ranging between 0, 2, 3 and 7 bp, with preference for TA or TNA containing targets (Chandler and Mahillon 2002; www-is.biotoul.fr). There are two known eukaryotic DNA TE superfamilies which are related to the IS5 group of bacterial TEs: Harbinger/PIF (containing Transposase_11 CD; 3-bp TSD) and IS4EU/ISL2EU (yielding no transposase-related CD-hits; 2-bp TSD) (Zhang *et al*. 2004; Kapitonov and Jurka 2004, 2007), which form distinct eukaryotic clades in Fig.3. The *A. vaga* IS5-like element, however, does not fall into any of these clades, and neither does it contain any additional ORFs, which represent one of the defining characteristics of Harbinger/PIF and IS4EU/ISL2EU superfamilies. Interestingly, all members of the ISL2 subgroup from the IS Finder database fall into two clades, rather than into a single ISL2 clade. We therefore designated the clade not containing ISL2 as the IS493 group, by the name of its first described representative.

Several IS5-like (Transposase_11 CD-containing) ORFs from *Trichomonas vaginalis* and certain other protists (*Entamoeba histolytica, Phytophthora spp., Aphanomyces euteiches*, *Hyaloperonospora parasitica*), instead of grouping with other eukaryotic clades, apparently form their own clade, albeit poorly supported (Fig. 3). One of *T. vaginalis* ORFs (TVAG_485520) is single-copy and is not framed by TIRs, suggesting domestication of this transposase-derived gene, while four others represent typical DNA TEs with a single ORF and 10-20 highly similar copies per *T. vaginalis* genome. In the *Trichomonas* families containing TVAG_135750 and TVAG_517480, the transposase domain is fused to the ubiquitin hydrolase-like cysteine peptidase (clan CA, family C19), while in the families containing TVAG_148970 and TVAG_413280, the transposase domain is not fused to any other domains. Despite the fact that these two groups of *Trichomonas* IS5-like families differ substantially in their amino acid sequence and their monophyly is not well supported, they share a peculiar feature: their relatively long TIRs (180-290 bp) contain shorter imperfect hairpin regions, ultimately resulting in formation of an imperfect direct repeat embedded into each of the inverted repeats. It is conceivable that this entire group (designated ISL2PR) evolved upon invasion of the ancestral protistan genome by an IS5-like transposon. However, IS5_Av does not fall into this group either. Instead, it occupies a very basal position in the ISL2 clade, together with two IS5-derived genes from the heterolobosean *Naegleria gruberi*. Since this clade is not well-supported, it is difficult to say whether an IS5-like element was transferred from bacteria into both *A. vaga* and *N. gruberi*, or between a *Naegleria*-like protist and *A. vaga*.

Expression analysis

Although the transposase-encoding IS5_Av ORF contains no in-frame stop codons or frameshifts that might indicate its non-functionality, it does exhibit an apparent deficiency: the first methionine is found at a position which would yield a protein that is 60-100 aa shorter than other transposases of the IS5 superfamily, and would not include the first region of similarity shared between all transposases at the N-terminus. As the use of splicing or ribosomal

frameshifting could in principle lead to production of a full-length transposase in the apparent absence of a correctly positioned ATG codon, we decided to examine the transcriptional activity of the element. RT-PCR analysis of *A. vaga* RNA (Fig. 2B) demonstrates that the level of IS5_Av transcripts, if present, is below the detection limits of the technique, while the positive control, the *A. vaga Ddl* gene (Gladyshev *et al*. 2008), as expected, yields a band corresponding to a transcribed and spliced message. The lack of transcriptional activity may account for IS5 inability to give rise to additional copies, perhaps due to the incompatibility of the promoter sequences with the host transcriptional machinery. Alternatively, the element may have been inactivated by a deletion interfering with RNA stability. Finally, IS5_Av may have been inactivated by a frameshift due to a replication slippage in a T8 stretch at pos. 782, whereby deletion of a T would yield a 317-aa ORF, adding 7 aa to the uninterrupted 310-aa polypeptide sequence, or addition of a T would yield a 373-aa ORF. Such an extended ORF, however, does not exhibit additional similarity to the N-termini of any known elements, and there is no apparent reason for transcriptional inactivation due to a single frameshift.

We also considered the possibility that IS5_Av may lack an appropriately positioned ATG codon because of utilization of an alternative start codon in the previous host. To check the likelihood of this scenario, we attempted to evaluate the possible usage of alternative initiation codons based on gene prediction models. We scanned the IS5_Av sequence with EasyGene (Larsen and Krogh 2003; Nielsen and Krogh 2005), which uses a high-quality training set of genes coding for known conserved proteins from each genome to estimate HMM (hidden Markov models) of gene prediction for that particular genome. Of the 138 species with HMM models in the database, 48 did not yield a predicted gene in the 2020-bp IS5_Av sequence; the first ATG codon at pos. 1040 was predicted as the optimal start site in 46 species, and as a suboptimal start site in 5 species; the TTG codon at pos. 788 (yielding a 308-aa transposase) was found to be optimal in 24 species and suboptimal in 2 species; and the GTG codon at pos. 884 (yielding a 276-aa transposase) was found to be optimal in 3 species and suboptimal in 2

species (see Table 3 for a list of alternative codons, and Table S1 for a complete list of predicted start codons). Thus, for approximately one-third of all database species that yielded gene predictions, the alternative start codons were predicted to be optimal. While we do not yet know the identity of the putative donor species, it is worth noting that the genera such as Bacteroides, Burkholderia and Pseudomonas, all of which are represented in Table 3, have already been identified as putative donors of foreign genes to bdelloids (Gladyshev *et al*. 2008).

Discussion

In this study, we describe an unusual DNA transposon IS5_Av from the DDE megafamily of transposases/integrases, which was found in the genome of a multicellular animal, but appears more similar to prokaryotic than to eukaryotic counterparts. It is framed by perfect 213-bp terminal inverted repeats and contains a single ORF coding for an IS5-like transposase. Interestingly, the host contains only a single copy of this element per diploid genome. Single-copy TEs are quite rare in eukaryotes, and whenever one TE copy per genome is reported, it is usually identified in a search of genome databases, which are far from being complete, especially when it comes to repetitive regions that are often left unassembled and are missing from most databases. We, however, verified the single-copy status of IS5_Av by Southern blot hybridization and library screening, and are confident that no other copy of this element is present elsewhere in the *A. vaga* genome.

In a standard life cycle of DNA TEs, there are two stages with a single-copy status (Fig. 4): (i) at the time of entry of a single invading copy, and (ii) just before elimination of the TE from the genome, when all of its copies but the last one have already been lost. (The third possibility, *i.e*. indefinite maintenance of a single domesticated copy in the genome by purifying selection due to acquisition of a cellular function, usually involves loss of terminal inverted repeats and is therefore highly unlikely.) At which of these two time points did we find IS5_Av? If it were about to become lost from the *A. vaga* genome after having peaked in copy number, it may be

expected to have undergone extensive ORF-damaging mutational decay, often accompanied by secondary insertions, so that the coding region would require molecular reconstruction and the terminal inverted repeats would have accumulated differences. The perfect identity of the TIRs and the absence of interruptions in the ORF make the recent entry hypothesis more plausible. The lack of expression from a single genomic copy is indicative of its pseudogene nature, and the integrity of a ~300-aa pseudogene would be typically compromised by mutational decay in a few million years (Lynch and Conery 2000, 2003). IS5_Av resides within an ancient *hobo*-like element, which underwent insertions of two TEs and, in addition, carries defects in its ORF (a frameshift and an in-frame stop codon). While the *hobo* itself is rather decayed, IS5_Av is most likely inserted into the 3' UTR of *hobo*, since the distance between the *hobo* stop codon and the IS5_Av insertion site is only 108 bp and does not include polyadenylation signals (which are found downstream from the IS5_Av insertion). A search for TSD-TIR for *hobo* revealed only one putative 8bp-TIR-TIR-8bp combination, fully consistent with IS5_Av insertion into the *hobo* 3' UTR. Thus, it appears that IS5_Av underwent integration into the *A. vaga* chromosome on its own, and not as a component of delivery vehicles such as phages or viruses, fragments of which would have been detectable in the adjacent genomic environment (but see further discussion below).

Naturally occurring recent inter-kingdom movements of TEs have not yet been reported in the literature, and our finding could represent a rare example of HGT between bacteria (or protists) and multicelluar animals, in which the TE is apparently "caught in the act" of transfer at the time when it failed to increase in copy number. Although the putative donor species is yet to be identified, and there are no other IS5_Av copies in the genome which could be used to determine the level of divergence between copies so as to estimate their arrival time, this HGT event may be regarded as relatively recent for reasons discussed above. The absence of detectable transcription from IS5_Av raises questions regarding the mechanisms responsible for its transfer and successful one-time integration into the *A. vaga* genome. It may be thought that

DNA TEs are more prone to HGT because they can be transmitted as transpososomes – nucleoprotein complexes containing TE DNA and the element-encoded transposase, which are sufficient for integration into the target *in vitro* (in the presence of Mg^{2+}) as well as *in vivo*. This capacity of the bacterial Tn5 transpososome, for instance, was even utilized in commercial applications (Reznikoff *et al*. 2004). Perhaps HGT of IS5_Av into *A. vaga* occurred when the transposon entered the germ line *via* routes operating during HGT of other foreign genes (Gladyshev *et al*. 2008), or possibly as a complex with the element-encoded transposase, which was able to complete the integration reaction. However, subsequent adaptation of the transcription/translation control sequences to the new host did not take place, and therefore its genomic mobility was restricted to the initial integration event, since it has never had a chance to produce transposase molecules in the new host, or perhaps yielded only misfolded and/or mislocalized molecules, even if the level of transcription (undetectable by PCR) was sufficient to produce any.

The much less likely “post-amplification extinction” scenario would involve loss of the functional incoming copy plus any other transposed copies, including incomplete remnants, and retention of a single non-functional transposed copy, which underwent loss of the ATG codon *via* mutation or deletion. However, mutation of the ATG codon *per se* does not imply an immediate loss of promoter activity and cessation of transcription. The absence of a correctly positioned ATG codon could rather indicate that IS5_Av was using a non-canonical initiation codon, such as TTG or GTG, in its previous host. It is also formally possible that an ATG-less incoming DNA copy was bound to a functional transposase produced *in trans*, and entered the *A. vaga* genome as a transposition complex for one-time integration.

Another imaginable scenario would combine recent invasion with nearly-immediate loss: the element could have arrived to a different telomere, already embedded in foreign sequences, *via* the same pathways which permit overall acquisition of foreign DNA (Gladyshev *et al*. 2008), but, because of the dynamic nature of bdelloid telomeres, the original invading copy, with these

adjacent sequences, was lost upon chromosome end erosion, and what we are seeing is a daughter copy that underwent a round of transposition in *A. vaga*, but lost its functionality either during or soon after transposition. This explanation assumes that the invading copy was fully capable of expression and transposition in *A. vaga*, but was lost very rapidly without a chance to spur more than one daughter copy.

Could IS5_Av, displaying a very basal phylogenetic position in the ISL2 clade, represent a highly diverged member of a eukaryotic IS4/IS5-like TE superfamily, such as Harbinger/PIF, IS4EU/ISL2EU, or ISL2PR, which, however, artificially clusters with ISL2-like transposons as a result of homoplasy? It should be emphasized that members of both previously described eukaryotic IS4/IS5-like superfamilies typically contain another ORF in addition to transposase: in case of Harbingers, this extra ORF is characterized by a SANT/Myb/trihelix motif (Kapitonov and Jurka 2004), while for IS4EU/ISL2EU it resembles a lambda DNA exonuclease, and the transposase itself contains an N-terminal DNA-binding THAP domain (Kapitonov and Jurka 2007). Since none of these extra ORFs (and domains) could be identified in IS5_Av, it appears likely that the element belongs to a single-ORF family, as do most other bacterial TEs related to IS4 and IS5 (www-is.biotoul.fr; Chandler and Mahillon 2002). The newly identified protistan IS5-like elements also differ substantially from IS5_Av, both by phylogenetic placement and by the TIR structure. Neither does IS5_Av represent a domesticated single-copy TE (as is the case for many Harbinger-derived genes), because such domestication usually involves loss of TIRs, which are present in IS5_Av and extend for 213 bp without a single mismatch. While bacterial TIRs in the ISL2 group usually fall within 10-40 bp range (http://www-is.biotoul.fr), exceptionally longer TIRs (up to 214 bp) are rare but not unprecedented in other bacterial IS groups (Kholodii *et al*. 2000).

Finally, one may entertain a possibility that this deep-branching element is the only known member of a novel eukaryotic IS5-related DNA TE superfamily, which is so rare that its other members, if any, have not yet been identified in any of the eukaryotic genomes

sequenced to date. Future projects aimed at sequencing previously unexplored eukaryotic genomes may be able to supply us with new representatives of this hypothetical superfamily.

Supplementary Material

Supplementary Data File: Multiple sequence alignment of IS5-like transposases.

Supplementary Table: Initiation codons in IS5_Av predicted by EasyGene.

Acknowledgments

We wish to thank Bill Reznikoff and two anonymous reviewers for valuable comments and suggestions. This research was supported by the U.S. National Science Foundation grant MCB-0821956 to I.A.

Literature Cited

Arkhipova IR, Meselson M. 2005 Diverse DNA transposons in rotifers of the class Bdelloidea. Proc Natl Acad Sci USA. 10233:11781-11786.

Chandler M, Mahillon J. 2002 Insertion sequences revisited. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, eds. Mobile DNA II. Washington D.C.: American Society for Microbiology; pp. 305–366.

Daniels SB, Petterson KR, Strausbaugh LD, Kidwell MG, Chovnick AC. 1990. Evidence for horizontal transmission of the P transposable elements between Drosophila species. Genetics. 124:339–355.

Diao X, Freeling M, Lisch D. 2006. Horizontal transfer of a plant transposon. PLoS Biol. 41:e5.

Gualerzi, C.O, Pon CL. 1990. Initiation of mRNA translation in prokaryotes. Biochemistry. 29:5881–5889.

Gladyshev EA, Arkhipova IR. 2007. Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci USA. 104:9352-9357.

Gladyshev EA, Meselson M, Arkhipova IR. 2007. A deep-branching clade of retrovirus-like retrotransposons in bdelloid rotifers. Gene. 390:136-145.

Gladyshev EA, Meselson M, Arkhipova IR. 2008. Massive horizontal gene transfer in bdelloid rotifers. Science. 320:1210-1213.

Gueiros-Filho FJ, Beverley SM. 1997. Trans-kingdom transposition of the Drosophila element mariner within the protozoan Leishmania. Science. 276:1716-1719.

Hartl DL, Lozovskaya ER, Nurminsky DI, Lohe AR. 1997. What restricts the activity of mariner-like transposable elements. Trends Genet. 135:197-201.

Kapitonov VV, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. DNA Cell Biol. 235:311-324.

Kapitonov VV, Jurka J. 2007. IS4EU, a novel superfamily of eukaryotic DNA transposons. Repbase Reports. 74:143-147.

Kidwell MG. 1983. Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. Proc Natl Acad Sci USA. 80:1655–1659.

Kholodii G, Yurieva O, Mindlin S, Gorlenko Z, Rybochkin V, Nikiforov V. 2000. Tn5044, a novel Tn3 family transposon coding for temperature-sensitive mercury resistance. Res Microbiol. 1514:291-302.

Larsen TS, Krogh A. 2003. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics. 4:21.

Lynch M, Conery JS. 2000 The evolutionary fate and consequences of duplicate genes. Science. 290:1151-1155.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 31:35-44.

Nielsen P, Krogh A. 2005. Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics. 21:4322-4329.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205-217.

Pace JK, Gilbert C, Clark MS, Feschotte C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci USA. 105:17023-17028.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 317:86-94.

Reeve JN. 1993. Structure and organization of genes, p. 493–527. In JG Ferry, ed., Methanogenesis: ecology, physiology, biochemistry and genetics. Chapman and Hall, New York.

Reznikoff WS, Goryshin IY, Jendrisak JJ. 2004.Tn5 as a molecular genetics tool: In vitro transposition and the coupling of in vitro technologies with in vivo transposition. Methods Mol Biol. 260:83-96.

Robertson HM. 1993. The mariner transposable element is widespread in insects. Nature. 362:241-245.

Robertson HM, Lampe DJ. 1995. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. Mol Biol Evol. 125:850-862.

Rubin EJ, Akerley BJ, Novik VN, Lampe DJ, Husson RN, Mekalanos JJ. 1999. *In vivo* transposition of mariner-based elements in enteric bacteria and mycobacteria. Proc Natl Acad Sci USA. 964:1645-1650.

Rubin E, Lithwick G, Levy AA. 2001. Structure and evolution of the hAT transposon superfamily. Genetics. 158:949-957.

Silva JC, Loreto EL, Clark JB. 2004. Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol. 6:57–72.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596-1599.

Zhang JK, Pritchett MA, Lampe DJ, Robertson HM, Metcalf WW. 2000. *In vivo* transposon mutagenesis of the methanogenic archaeon *Methanosarcina acetivorans* C2A using a modified version of the insect mariner-family transposable element Himar1. Proc Natl Acad Sci USA. 97:9665-9670.

Zhang X, Jiang N, Feschotte C, Wessler SR. 2004. PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. Genetics. 166:971-986.

Figure Legends

FIG. 1.–Structure (A) and genomic environment (B) of IS5_Av. (**A**) Within the transposase (TPase) ORF, the positions 788, 884, 1040, and 1714 denote the TTG, GTG, ATG, and TGA codons, respectively. Also shown are the positions of the DDE catalytic residues, and the primers used for PCR amplification of the internal region between TIRs and for RT-PCR shown in Fig. 2 (RT-F1, RT-R1) (half-arrows). A small gray box denotes a putative 63-aa ORF extension in case of a -1 frameshift. Hypothetical 10-bp secondary TIRs are shown by tiny arrowheads. Primers for the outermost part of TIRs (unnamed half-arrows) failed to yield PCR products, apparently due to PCR interference with the DNA secondary structure. Scale bar, 100 bp. (**B**) A 40-kb region (region 63135..103135 from GenBank accession No. EU643477) near the *A. vaga* telomere K_A (designated by the letter T), with nine TE insertions (Table 2) and five protein-coding ORFs, of which histidine-ammonia lyase (HAL, 3' truncated) is presumably of bacterial origin, acyl-CoA synthetase (ACS, 3' truncated) is of metazoan origin, and the rest is of uncertain origin. Coding sequences with transcriptional orientation towards the telomere are shown above the line, while those with transcriptional orientation away from the telomere are shown below the line. Two AATAAA signals in the *hobo* 3' UTR are denoted by letters A. The intergenic region scanned for TSD-TIR combinations is denoted by a thick line. Scale bar, 1 kb.

FIG. 2.–Copy number (A) and expression (B) of IS5_Av. (**A**) Ethidium bromide staining (left panels) and Southern blot hybridization with the ^{32}P-labeled IS5_Av probe (right panels) of the *A. vaga* genomic DNA digested with *Pvu*II or *Xho*I/*Hpa*I, as indicated. (**B**) RT-PCR with IS5_Av (left) and Ddl_Av (right) forward and reverse primers. Shown are the sizes of the corresponding PCR products, which for the *Ddl* gene differ by 50 bp due to splicing. RT-, no reverse transcriptase added; RT+, addition of SuperScriptII; DNA, control genomic DNA amplification. M, 1 kb+ ladder (Invitrogen).

FIG. 3.–Phylogenetic placement of IS5_Av. Shown is a neighbor-joining phylogram including representatives of different groups of the IS5 megafamily, which contains six known groups of the bacterial IS5 family (IS5, IS903, ISH1, IS1031, IS427, ISL2) and two known eukaryotic IS5-like superfamilies (Harbinger/PIF and IS4EU/ISL2EU). Also shown are the newly designated groups ISL2PR (from various protists) and IS493 (members of which are assigned to ISL2 in the IS Finder database). Bootstrap support values for the major named groups, obtained by neighbor joining and minimum evolution methods, respectively, are as follows: ISL2, 48/58; IS493, 90/89; IS5, 100/100; IS903, 100/100; ISH1, 100/100; IS1031, 100/100; IS427, 100/100; ISL2PR, 34/43; IS4EU/ISL2EU, 100/100; Harbinger/PIF, 51/47. The bacterial IS5, IS903, ISH1, IS1031, and IS427 groups can be joined into a well supported (97/87) supergroup. Support for IS5_Av/Naegleria branch is 51/64. Eukaryotic species are in boldface, and branches leading to them are shown by thick lines. Scale bar, 0.1 amino acid substitutions per site. For multiple sequence alignment, see Supplementary Data.

FIG. 4.–Copy number dynamics of DNA TEs (modified after Hartl *et al*. 1997). The shape of the curve is arbitrary, as is the location of branches giving rise to domesticated or horizontally escaped copies.

Table 1. BLASTP similarity scores for the IS5_Av query used to search the non-redundant GenBank database (A) and the IS finder database (http://www-is.biotoul.fr/) (B). Two IS5-derived ORFs from *Naegleria gruberi* yield the same top hits as IS5_Av. Note that most of the GenBank entries are misannotated as IS4 transposases.

(A)Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
7,031,513 sequences; 2,427,977,331 total letters
Query=IS5_Av
Length=310

Sequences producing significant alignments:	Score (Bits)	E Value
ref|YP_594226.1| transposase, IS4 [Deinococcus geothermalis D...	81.3	1e-13
ref|ZP_02737947.1| transposase, IS4 [Gemmata obscuriglobus UQ...	77.8	1e-12
ref|ZP_02732909.1| transposase, IS4 [Gemmata obscuriglobus UQ...	76.6	2e-12
ref|ZP_03176005.1| putative transposase [Streptomyces sp. SPB...	75.1	7e-12
ref|ZP_03143215.1| transposase IS4 family protein [Cyanothece...	70.5	1e-10
ref|ZP_03141860.1| transposase IS4 family protein [Cyanothece...	69.3	3e-10
ref|ZP_03178672.1| putative transposase [Streptomyces sp. SPB...	68.6	6e-10
ref|ZP_03143523.1| transposase IS4 family protein [Cyanothece...	68.6	6e-10
ref|ZP_02942299.1| transposase IS4 family protein [Cyanothece...	68.6	7e-10
ref|ZP_02939447.1| transposase IS4 family protein [Cyanothece...	68.2	7e-10
ref|ZP_01731099.1| hypothetical protein CY0110_01550 [Cyanoth...	68.2	9e-10
ref|ZP_02942406.1| transposase IS4 family protein [Cyanothece...	67.4	1e-09

(B) Database: IS protein Database
3402 sequences; 1,042,201 total letters
Query= IS5_Av
(310 letters)

Sequences producing significant alignments	IS Family	Group	Origin	Score (bits)	E (value)
ISDge6	IS5	ISL2	Deinococcus geothermalis	77	1e-15
IS1515	IS5	ISL2	Streptococcus pneumoniae I41R	62	3e-11
IS702	IS5	ISL2	Calothrix sp. PCC7601	62	4e-11
ISL2A	IS5	ISL2	Lactobacillus helveticus LH27	54	1e-08
ISL2	IS5	ISL2	Lactobacillus helveticus LH28	54	1e-08
ISMae4	IS5	ISL2	Microcystis aeruginosa	53	2e-08
IS493	IS5	ISL2	Streptomyces lividans CT2	45	5e-06
IS1381	IS5	ISL2	Streptococcus pneumoniae	42	4e-05
IS470	IS5	ISL2	Streptomyces coelicolor A3(2) M145	42	5e-05
IS1373	IS5	ISL2	Streptomyces lividans 66 1326.32	40	2e-04
IS1381A [V]	IS5	ISL2	Streptococcus agalactiae A909	39	3e-04
IS112	IS5	ISL2	Streptomyces albus G J1147	39	4e-04

Table 2. Top BLASTP database hits for eight TEs from the *A. vaga* telomere K_A (EU643477) shown in Fig. 1B. Also listed are hits to the conserved domain (CD) database, and disruptions in ORFs, if any.

TE superfamily (position)	Top BLASTP hit, species	E-value	% identity (similarity)	CD-hits	In-frame stops/ frameshifts/indels
mariner (66428..68107)	Transposase, *Pachygrapsus marmoratus* (coastal crab)	5e-61	36% (54%)	Transposase_1	2 / 1 / 0
piggyBac (68182..70303)	Transposase, *Acyrthosiphon pisum* (pea aphid)	9e-89	42% (63%)	none	1 / 0 / 0
mariner (71477..71555, 72009..72645)	Avmar1 transposase, *Adineta vaga* (bdelloid rotifer)	1e-78	78% (83%)	Transposase_1	1 / 1 / 2
hobo (77307..78830, 80417..81399)	Transposase, *Bactrocera tryoni* (Queensland fruit fly)	2e-39	24% (44%)	hATC superfamily	1 / 1 / 1
mariner (78833..80416)	Transposase, *Caenorhabditis elegans* (roundworm)	2e-40	31% (50%)	none	0 / 0 / 0
IS5 (81508..83527)	Transposase, *Deinococcus geothermalis* (radio-resistant micrococci)	1e-13	25% (44%)	Transposase_11	0 / 0 / 0
LTR retrotransposon (*gag; pol; env*) (84198..93163)	Retrotransposon *gag* protein, *Asparagus officinalis* (Liliopsida); *pol* polyprotein, *Danio rerio* (zebrafish); *env*-like, transmembrane glycoprotein, coronavirus	1e-06; 0.0; 3e-01	21%(44%); 39% (57%) 27% (47%)	Retrotrans_gag; RVP superfamily; RT_LTR (RT-like superfamily); rve superfamily	0 / 0 / 0; 0 / 2 / 0; 0 / 0 / 0
piggyBac (100845..102137)	Transposase, *Nasonia vitripennis* (jewel wasp)	6e-18	31% (54%)	none	6 / 0 / 1

Table 3. List of putative donor species predicted by EasyGene to yield a full-length transposase using an alternative start codon as an optimal. Four species predicted to utilize alternative start codons as suboptimal (CDSsub) are also included.

Model	Feature	Start	End	Score	Startc	Species	Taxonomy
SSW02	CDS	788	1714	8.60E-27	TTG	*Synechococcus sp. WH 8102*	Cyanobacteria; Chroococcales
PMMI02	CDS	788	1714	9.81E-22	TTG	*Prochlorococcus marinus str. MIT 9313*	Cyanobacteria; Prochlorales
NM02	CDS	884	1714	3.96E-11	GTG	*Neisseria meningitidis serogroup B MC58*	Proteobacteria; Betaproteobacteria; Neisseriales
SE02	CDS	788	1714	7.16E-11	TTG	*Salmonella enterica subsp. enterica*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
NO02	CDS	788	1714	1.78E-10	TTG	*Nostoc sp. PCC 7120*	Cyanobacteria; Nostocales
PL01	CDS	788	1714	4.16E-10	TTG	*Photorhabdus luminescens subsp. laumondii*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
LP02	CDS	788	1714	5.54E-09	TTG	*Lactobacillus plantarum WCFS1*	Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae
ECC02	CDS	788	1714	1.66E-08	TTG	*Escherichia coli CFT073*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
STT02	CDS	788	1714	1.93E-08	TTG	*Salmonella enterica subsp. enterica*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
STY02	CDS	788	1714	4.07E-08	TTG	*Salmonella typhimurium LT2*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
ECE03	CDS	788	1714	8.15E-08	TTG	*Escherichia coli O157:H7 EDL933*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
WDM01	CDS	788	1714	1.12E-07	TTG	*Wolbachia endosymbiont of D. melanogaster*	Proteobacteria; Alphaproteobacteria; Rickettsiales
LM02	CDSsub	884	1714	1.21E-07	GTG	*Listeria monocytogenes EGD*	Firmicutes; Bacilli; Bacillales
CDI01	CDS	788	1714	1.54E-07	TTG	*Corynebacterium diphtheriae*	Actinobacteria; Actinobacteridae; Actinomycetales
BS03	CDS	788	1714	4.32E-07	TTG	*Bacillus subtilis*	Firmicutes; Bacilli; Bacillales
CT02	CDS	788	1714	4.96E-07	TTG	*Chlamydia trachomatis*	Chlamydiae/Verrucomicrobia; Chlamydiae; Chlamydiales
ECO02	CDS	788	1714	5.17E-07	TTG	*Escherichia coli O157:H7*	Proteobacteria; Gammaproteobacteria; Enterobacteriales
XC02	CDS	788	1714	6.53E-07	TTG	*Xanthomonas campestris pv. campestris*	Proteobacteria; Gammaproteobacteria; Xanthomonadales
BT02	CDS	788	1714	8.86E-07	TTG	*Bacteroides thetaiotaomicron VPI-5482*	Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidales
VC02	CDS	788	1714	1.41E-06	TTG	*Vibrio cholerae*	Proteobacteria; Gammaproteobacteria; Vibrionales
BS03	CDSsub	884	1714	1.49E-06	GTG	*Bacillus subtilis*	Firmicutes; Bacilli; Bacillales
CB02	CDSsub	788	1714	2.72E-06	TTG	*Coxiella burnetii RSA 493*	Proteobacteria; Gammaproteobacteria; Legionellales
NE02	CDS	788	1714	3.23E-06	TTG	*Nitrosomonas europaea*	Proteobacteria; Betaproteobacteria; Nitrosomonadales
MBU01	CDS	788	1714	7.20E-06	TTG	*Methanococcoides burtonii DSM 6242*	Euryarchaeota; Methanomicrobia; Methanosarcinales
CB02	CDS	884	1714	7.76E-06	GTG	*Coxiella burnetii RSA 493*	Proteobacteria; Gammaproteobacteria; Legionellales
SO02	CDS	788	1714	8.70E-06	TTG	*Shewanella oneidensis MR-1*	Proteobacteria; Gammaproteobacteria; Alteromonadales
XF02	CDSsub	788	1714	1.83E-05	TTG	*Xylella fastidiosa*	Proteobacteria; Gammaproteobacteria; Xanthomonadales
BPS01	CDS	788	1714	0.000366	TTG	*Burkholderia pseudomallei K96243*	Proteobacteria; Betaproteobacteria; Burkholderiales
PS02	CDS	788	1714	0.000449	TTG	*Pseudomonas syringae pv. tomato*	Proteobacteria; Gammaproteobacteria; Pseudomonadales
BBA01	CDS	884	1714	0.000589	GTG	*Bdellovibrio bacteriovorus*	Proteobacteria; Deltaproteobacteria; Bdellovibrionales
MLO03	CDS	788	1714	0.170785	TTG	*Mesorhizobium loti*	Proteobacteria; Alphaproteobacteria; Rhizobiales

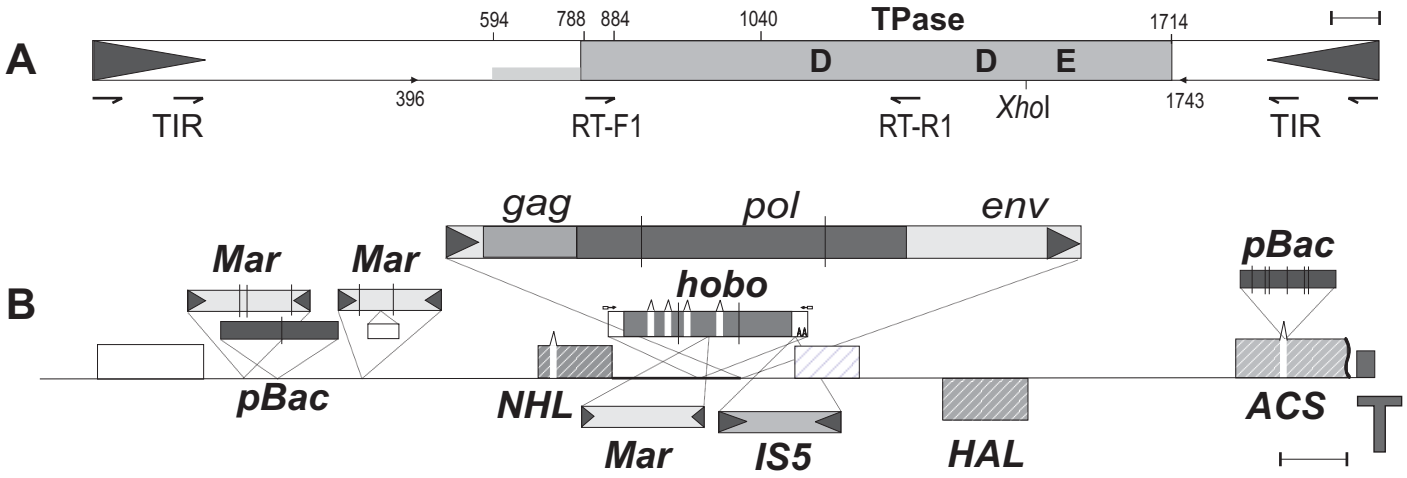
A
594
788
884
1040
TPase
1714
D
D
E
396
XhoI
1743
TIR
RT-F1
RT-R1
TIR
B
gag
pol
env
Mar
Mar
hobo
pBac
pBac
NHL
Mar
IS5
HAL
ACS

Fig. 1

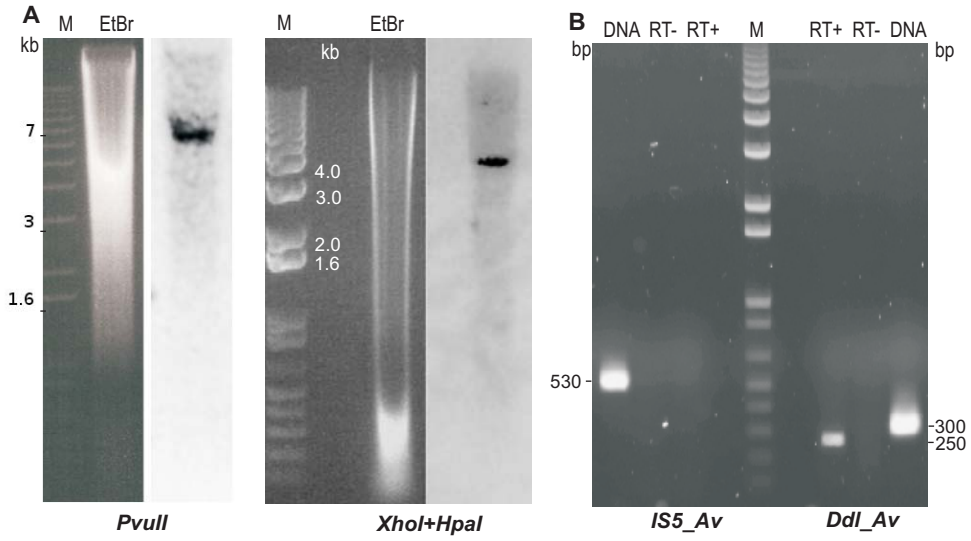
A
M
EtBr
kb
7
3
1.6
PvuII
M
EtBr
kb
4.0
3.0
2.0
1.6
XhoI+HpaI
B
DNA
RT-
RT+
M
RT+
RT-
DNA
bp
bp
530
300
250
IS5_Av
Ddl_Av

Fig. 2

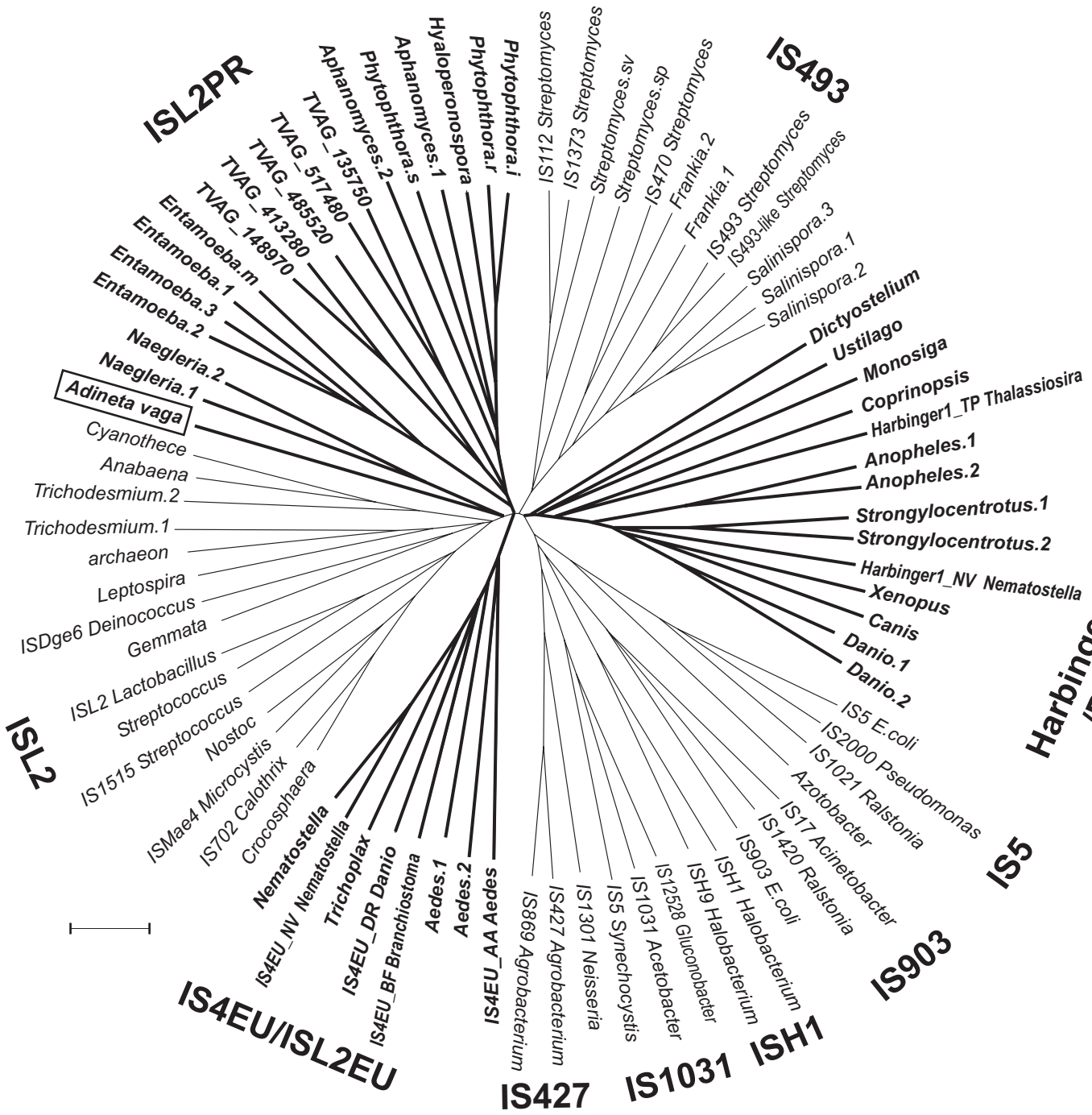
ISL2PR
IS493
Harbinger /PIF
IS5
IS903
ISH1
IS1031
IS427
IS4EU/ISL2EU
ISL2
Phytophthora.i
Phytophthora.r
Hyaloperonospora
Aphanomyces.1
Phytophthora.s
Aphanomyces.2
TVAG_135750
TVAG_517480
TVAG_485520
TVAG_413280
TVAG_148970
Entamoeba.m
Entamoeba.1
Entamoeba.3
Entamoeba.2
Naegleria.2
Naegleria.1
Adineta vaga
Cyanothece
Anabaena
Trichodesmium.2
Trichodesmium.1
archaeon
Leptospira
ISDge6 Deinococcus
Gemmata
ISL2 Lactobacillus
Streptococcus
IS1515 Streptococcus
Nostoc
ISMae4 Microcystis
IS702 Calothrix
Crocosphaera
Nematostella
IS4EU_NV Nematostella
Trichoplax
IS4EU_DR Danio
IS4EU_BF Branchiostoma
Aedes.1
Aedes.2
IS4EU_AA Aedes
IS869 Agrobacterium
IS427 Agrobacterium
IS1301 Neisseria
IS5 Synechocystis
IS1031 Acetobacter
IS12528 Gluconobacter
ISH9 Halobacterium
ISH1 Halobacterium
IS903 E.coli
IS1420 Ralstonia
IS17 Acinetobacter
Azotobacter
IS1021 Ralstonia
IS2000 Pseudomonas
IS5 E.coli
Danio.2
Danio.1
Canis
Xenopus
Harbinger1_NV Nematostella
Strongylocentrotus.2
Strongylocentrotus.1
Anopheles.2
Anopheles.1
Harbinger1_TP Thalassiosira
Coprinopsis
Monosiga
Ustilago
Dictyostelium
Salinispora.2
Salinispora.1
Salinispora.3
IS493-like Streptomyces
IS493 Streptomyces
Frankia.2
Frankia.1
IS470 Streptomyces
Streptomyces.sp
Streptomyces.sv
IS1373 Streptomyces
IS1122 Streptomyces

Fig. 3

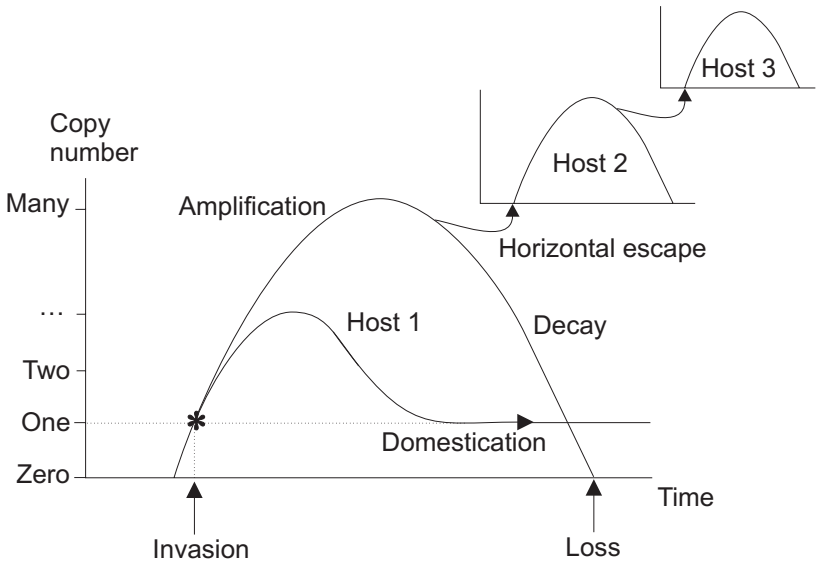
Host 3
Copy
number
Host 2
Many
Amplification
Horizontal escape
...
Host 1
Decay
Two
One
Domestication
Zero
Time
Invasion
Loss

Fig. 4